

Generating synthetic data in finance: opportunities, challenges and pitfalls

Year: 2020 | Citations: 262 | Authors: Samuel A. Assefa

Abstract

Financial services generate a huge volume of data that is extremely complex and varied. These datasets are often stored in silos within organisations for various reasons, including but not limited to regulatory requirements and business needs. As a result, data sharing within different lines of business as well as outside of the organisation (e.g. to the research community) is severely limited. It is therefore critical to investigate methods for synthesising financial datasets that follow the same properties of the real data while respecting the need for privacy of the parties involved. This introductory paper aims to highlight the growing need for effective synthetic data generation in the financial domain. We highlight three main areas of focus that are of particular importance while generating synthetic financial datasets: 1) Generating realistic synthetic datasets. 2) Measuring the similarities between real and generated datasets. 3) Ensuring the generative process satisfies any privacy constraints. Although these challenges are also present in other domains, the additional regulatory and privacy requirements within financial services present unique questions that are not asked elsewhere. Due to the size and influence of the financial services industry, answering these questions has the potential for a great and lasting impact. Finally, we aim to develop a shared vocabulary and context for generating synthetic financial data using two types of financial datasets as examples.